Final Report for Period: 07/2006 - 06/2007 **Submitted on:** 07/16/2007
Principal Investigator: Park, Haesun . **Award ID:** 0549247
Organization: GA Tech Res Corp - GIT
Title:
ALGORITHMS: Collaborative Research: Development of Vector Space based Methods for Protein Structure Prediction

Project Participants

Senior Personnel

Name: Park, Haesun
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Name: Kim, Hyunsoo
Worked for more than 160 Hours: Yes
Contribution to Project:

Graduate Student

Name: Lee, Sangmin
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Howland, Peg
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Park, Cheonghee
Worked for more than 160 Hours: Yes
Contribution to Project:

Undergraduate Student

Technician, Programmer

Name: Drake, Barry
Worked for more than 160 Hours: Yes
Contribution to Project:

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Dr. Chris Ding
Prof. Lars Elden
Prof. Gene Golub

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Findings: (See PDF version submitted by PI at the end of the report)

Training and Development:

Outreach Activities:

Journal Publications

J. Ye, R. Janardan, Q. Li, and H. Park, "Feature extraction via generalized uncorrelated linear discriminant analysis", IEEE Transactions on pattern analysis and machine intelligence, p. 1312, vol. 18-10, (2006). Published,

J. Ye, Q. Li, H. Xiong, R. Janardan, V. Kumar, and H. Park, "IDR/QR: An incremental dimension reduction algorithm via QR decomposition", IEEE Transaction on Knowledge and Data Engineering, Special Issue- Intelligent Data Preparation, p. 1208, vol. 17-9, (2005). Published,

C. Park, M. Jeon, P. Pardalos, and H. Park, "Quality assessment of gene selection in microarray data", Optimization methods and software, p. 145, vol. 22-1, (2007). Published,

H. Kim, B.L. Drake, and H. Park, "Adaptive nonlinear discriminant analysis by regularized minimum squared errors", IEEE Transactions on Knowledge and Data Engineering, p. 603, vol. 18-5, (2006). Published,

P. Howland, J. Wang, and H. Park, "Solving the small sample size problem in face recognition using generalized discriminant analysis", Pattern Recognition, p. 277, vol. 39-2, (2006). Published,

Park, CH; Park, H, "Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition", SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS, p. 87, vol. 27, (2005). Published, 10.1137/S089547980444233

Howland, P; Jeon, M; Park, H, "Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition", SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS, p. 165, vol. 25, (2003). Published,

Kim, H; Golub, GH; Park, H, "Missing value estimation for DNA microarray gene expression data: local least squares imputation", BIOINFORMATICS, p. 187, vol. 21, (2005). Published, 10.1093/bioinformatics/bth49

Kim, H; Howland, P; Park, H, "Dimension reduction in text classification with support vector machines", JOURNAL OF MACHINE LEARNING RESEARCH, p. 37, vol. 6, (2005). Published,

Kim, H; Park, H, "Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor", PROTEINS-STRUCTURE FUNCTION AND GENETICS, p. 557, vol. 54, (2004). Published, 10.1002/prot.1060

Howland, P; Park, H, "Generalizing discriminant analysis using the generalized singular value decomposition", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, p. 995, vol. 26, (2004). Published,

Ye, JP; Janardan, R; Park, CH; Park, H, "An optimization criterion for generalized discriminant analysis on undersampled problems", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, p. 982, vol. 26, (2004). Published,

Park, CH; Park, H, "Fingerprint classification using fast Fourier transform and nonlinear discriminant analysis", PATTERN RECOGNITION, p. 495, vol. 38, (2005). Published, 10.1016/j.patcog.2004.08.01

Park, CH; Park, H, "Nonlinear feature extraction based on centroids and kernel functions", PATTERN RECOGNITION, p. 801, vol. 37, (2004). Published, 10.1016/j.patcog.2003.07.01

Kim, H; Park, H, "Protein secondary structure prediction based on an improved support vector machines approach", PROTEIN ENGINEERING, p. 553, vol. 16, (2003). Published, 10.1093/protein/gzg07

C. Park and H. Park, "A relationship between LDA and the generalized minimum squared error solution", SIAM Journal on Matrix Analysis and Applications, p. 474, vol. 27-2, (2005). Published,

H.Kim, J. Zhou, H. Morse, and H. Park, "A three-stage framework for gene expression data analysis by L1 norm support vector refression", International Journal of Bioinformatics Research and Applications, p. 51-62, vol. 1-1, (2005). Published,

Books or Other One-time Publications

H. Kim and H. Park, "Adaptive kernel classifiers based on matrix decomposition updates
for biological data analysis,
", ( ). Book, Accepted
Editor(s): Y.-Q. Zhang and J.C. Rajapakse,
Collection: Machine learning in bioinformatics,
Bibliography: John Wiley and Sons

H. Kim and H. Park, "Extracting unrecognized gene relationships from biomedical literature via matrix factorizations using a prior knowledge of gene relationships", (2006). Conference Proceedings, Published
Editor(s): M. Song and Z. Obradovic
Collection: Proceedings of ACM First International Workshop on Text Mining in Bioinformatics
Bibliography: Nov. 2006, pp. 60-67

H. Kim and H. Park, "Missing value estimation for DNA microarray gene expression data by alternating least sqaures", (2006). Conference Proceedings, Published
Collection: Proceedings of the tenth annual international conference on research in computational molecular biology
Bibliography: Venice, Italy, April 2-5.

H. Park and L. Elden, "Matrix rank reductino for data analysis and feature extraction", (2005). Book, Published
Editor(s): E.H. Kontoghiorghes
Collection: Handbook of parallel computing and statistics
Bibliography: CRC Press

P. Howland and H. Park, "Cluster-preserving dimension reduction methods for efficient classification of text data", (2003). Book, Published
Editor(s): M. Berry
Collection: Survey of text mining
Bibliography: Springer-Verlag

H. Kim and H. Park, "One-sided non-negative matrix factorization and non-negative centroid dimension reduction for text classification", (2006). Conference Proceedings, Published
Editor(s): M.D. Castellanos and M.W. Berry
Collection: Proceedings of the workshop on text mining, the 6th SIAM international conference on data mining
Bibliography: SIAM

H. Kim and H. Park, "Discriminant analysis using nonnegative matrix factorization for nonparametric multiclass classification", (2006).

Conference Proceedings, Published
Collection: Proc. of the IEEE international conference on granular computing, Atlanta, GA, May 10-12
Bibliography: pp. 182-187

H. Kim and H. Park, "Two-dimensional concept vector machines based on an ionic interaction model", (2005). Conference Proceedings, Published
Collection: Proc. of the IEEE international conference on neural networks and brain, Beijing, China, Oct. 13-15
Bibliography: vol. 3, pp. 1991-1995

C. Park and H. Park, "A comparitive study of linear and nonlinear feature extraction methods", (2004). Conference Proceedings, Published
Collection: Proc. for the fourth IEEE international conference on data mining, Brighton, UK, Nov. 2004
Bibliography: pp. 495-498

H. Kim, G. Golub, and H. Park, "Imputation of missing value in DNA microarray gene expression data", (2004). Conference Proceedings, Published
Collection: Prof. of the IEEE computer society bioinformatics conference, Stanford, CA
Bibliography: pp. 572-573

H. Kim and H. Park, "Incremental and decremental least sqaures support vector machine and its application to drug design", (2004). Conference Proceedings, Published
Collection: IEEE computer society bioinformatics conference, Stanford, CA, Aug.
Bibliography: pp. 656-657

J. Ye, H. Xiong, R. Janardan, V. Kumar, and H. Park, "An incremental dimension reduction algorithm via QR decomposition", (2004). Conference Proceedings, Published
Collection: Proc. for the ACM SIGKDD confernece, Seattle, WA, Aug.
Bibliography: pp. 364-373

J. Ye, R. Janardan, Q. Li, and H. Park, "Feature extraction via generalized uncorrelated linear discriminant analysis", (2004). Conference Proceedings, Published
Collection: Proc. of the 21st international conference on machine learning, Banff, Alberta, Canada, July.
Bibliography: pp. 895-902

P. Howland and H. Park, "Equivalence of several two-stage methods for linear discriminant analysis", (2004). Conference Proceedings, Published
Collection: Proc. of the fourth SIAM international conference on data mining, Kissimmee, FL, April
Bibliography: pp. 69-77

H. Kim and H. Park, "Data reduction in support vectgor machines by kernelized ionic interactgion model", (2004). Conference Proceedings, Published
Collection: Proc. of the fourth SIAM international conference on data mining, Kissimmee, FL, April.
Bibliography: pp. 507-511

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

The relationship between protein sequence, structure and function is a complex one, and much more can be learned from structure than sequence. However, the cost of producing a one dimensional sequence is much less than producing a structure. While recent developments

towards high throughput structure determination will likely impact this balance, the situation will continue to hold true. For this reason there is much interest in contributing to the recognition of three dimensional structure from protein sequence. An efficient computer-based algorithm for determining the most probable structure from the sequence with very low sequence identity,
would be a great benefit in this task.

The method that we investigated here is intended to provide critical information for prediction of protein 3D structure from accurate prediction of protein secondary structure and relative solvent accessibility. We have developed a unique numerical representation of proteins that have been represented by character strings of alphabets, reflecting long-range interactions. To achieve our goal of finding proper numerical representation of proteins, we explored encoding schemes based on the statistical matrix which is generated by the product of the relative occurrences of the two amino acids for a given long range interaction pair and on specific characteristics of the residue such as hydrophobicity or sequence around the residue within some predetermined window. Though the local sequence environment is most significant for secondary structure, the long range interaction may play an important role in protein folding including cross-linking secondary structure.

For protein structure prediction, we designed knowledge-based prediction systems which are scalable to data, efficient and effective. For this, we developed kernel based nonlinear extensions of classification and dimension reduction methods such as KOC and KDA/GSVD. Our experimental results indicate that conventional linear classifiers do not perform well on protein structure prediction. Accordingly, we developed nonlinear classifiers and nonlinear dimension reduction and visualization based on our earlier work which is applicable to linear problems, extending ideas from reproducing kernel Hilbert space.

In midroarray analysis, the gene selection methods and adaptive dimension reduction methods that we investigated are intended to provide critical tool in designing practical methods in handling massive data sets of very high dimension. The adaptive dimension reduction method that we have designed is important in efficient prediction of protein 3D structure using the machine learning methods such as the support vector machines. The adaptive methods make it possible to utilize a very large amount of data points without repeating all stages of expensive computation of the solution. When combined with dimension reduction, it produces a powerfully fast algorithm that allows us to utilize more data with a fixed amount of storage and computation power. Our earlier experimental results indicate that conventional linear classifiers do not perform well on protein structure prediction. Accordingly, we developed adaptive nonlinear classifiers and adaptive nonlinear dimension reduction and visualization based on our earlier work extending ideas from reproducing kernel Hilbert space. Out adaptive methods are based on the mathematical relationship between the minimum squared error formulation and the linear discriminant analysis and its nonlinear extension gives further importance and practical value to the linear and nonlinear dimension reduction methods to protein structure prediction.

Gene expression microarray data sets often contain missing expression values. Robust missing value estimation methods are needed since many algorithms for gene expression analysis require a complete matrix of gene array values. The identification of discriminant genes is one of the most fundamental steps in microarray gene expression data analysis to confirm recent discovery in disease research or suggest new genes to be explored.

Gene expression data sets often contain missing values for various reasons. For example, the background and the signal may have similar intensities, the surface of the chip may not be planar, there may be dust on the slides, the probe may not be properly fixed on the chip or washed properly, the hybridization step may not work properly. There are several approaches for estimating the missing values. Recently, for missing value estimation, the singular value decomposition based method (SVDimpute) and weighted k-nearest neighbors imputation (KNNimpute) have been introduced. It has been shown that KNNimpute performs better on non-time series data or noisy time series data, while SVDimpute works well on time series data with low noise levels. Overall, the weighted k-nearest neighbors based imputation provides a more robust method for missing value estimation than the SVD based method. Recently, Bayesian principal component analysis (BPCA), which simultaneously estimates a probabilistic model and latent variables within the framework of Bayesian inference, has been successfully applied to missing value estimation problems. A fixed rank approximation algorithm (FRAA) using the singular value decomposition has been proposed. However, FRAA could not outperform KNNimpute even though it is more accurate than replacing missing values with 0's or with row means.

We developed a novel least squares based imputation method in the gene expression data for estimating missing values. Our local least squares based method (LLSimpute) represents a target gene that has missing values as a linear combination of similar genes. The similar genes are chosen by k-nearest neighbors or the concept of coherent genes that have large absolute values of Pearson correlation coefficients. The imputation method has been tested on different data sets and compared with KNNimpute and an estimation method based on BPCA. The LLSimpute method has already been widely used and studied by researchers in the community.

The gene selection problem involves finding a minimum subset of the original features that yields the best classification performance. Many of the approaches for feature selection have been in use. Correlation coefficients are used to rank features for evaluating how well an individual

feature contributes the discrimination between the diseased and normal states. After training a multivariate classifier such as Fisher's linear discriminant or support vector machines, we can find the features with the most information content by ranking the weights of the corresponding features. The results produced by these methods are often far from being optimal if we remove more than one feature at a time. Moreover, the classical Fisher's linear discriminant fails when the number of features is larger than the number of data points due to potential singularity problems of the scattering matrices. Recursive feature elimination (RFE) with support vector machines was proposed to overcome this problem. This is based on backward feature elimination and a wrapper method, which requires very high computational cost. Recently, regression approaches have emerged where classification is considered as a special case of regression. Ridge regression tends to retain all elements of the weight vector since it uses L_2 penalty. Least absolute shrinkage and the selector operator (LASSO) uses L_1 penalty instead so that it zeroes out all but an optimal feature subset. The originally solution of LASSO cannot be applied when the number of features (n) is greater than the number of samples (m).

In our research, we used L_1-norm support vector machines with recursive multiple feature elimination for gene selection for cancer classification, which is accurate and computationally efficient since it can remove several features at the same time for the identification of discriminant genes. Our research has greatly increased the performance of the missing value estimation and gene selection for cancer classification. All discovered methods have been rigorously tested computationally using the microarray data sets which are publicly available including those from the Stanford Microarray Database (SMD).

Contributions to Other Disciplines:

Contributions to the Other Disciplines of Science or Engineering

In the process of designing effective encoding schemes for proteins and efficient algorithms for protein structure prediction, we have developed the methods that can be applied to various microarray analysis data. They are the missing value estimation and gene selection in microarray analysis. Unlike in other applications such as text processing, facial recognition, or finger print classification, feature extraction cannot give the useful information in microarray analysis where the information on significance of specific genes is needed. By finding a relationship between feature extract and feature selection, we have designed a powerful gene selection method based on the linear discriminant analysis and also L1 norm based support vector machines.

The adaptive linear and nonlinear dimension reduction algorithms and the missing value estimation methods that we designed have broad applications in Science and Engineering. The adaptive linear and nonlinear dimension reduction algorithms have already been successfully applied to applications such as facial recognition, text data classification in information retrieval, finger print classification. The missing value estimation method is needed in many other application areas such as collaborative filtering, and in any data analysis of experiments where there can be missing data points. We are in the process of further applying our methods to and this will be one of the topics in the next stage of our research.

We are exploring the possibility of applying our algorithms developed for the protein structure prediction and microarray data analysis to network intrusion detection. Interestingly, there is some similarity among these problems due to the fact that the original data sets are non-numeric and good numeric encoding schemes are needed to utilize vector space based methods. In addition, clustering and classification play main roles in both problems for class prediction of unseen data items.

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

All of the papers published with the support of the grant
are available to view and download at 'http://www.cc.gatech.edu/~hpark'.

Contributions Beyond Science and Engineering:

Categories for which nothing is reported:

Organizational Partners

Activities and Findings: Any Training and Development

Activities and Findings: Any Outreach Activities

Any Web/Internet Site

Any Product
Contributions: To Any Human Resource Development
Contributions: To Any Beyond Science and Engineering

Major Findings Resulting from the Activities

The major findings resulting from the research activities include the discovery of optimal encoding schemes for amino acids for protein structure prediction, efficient implementation of the support vector machine classifier utilizing its relationship to the linear discriminant analysis, efficient adaptive classifier design which can efficiently find the new solution relating to the data sets that are changing, new dimension reduction method LDA/GSVD and its nonlinear extension, gene selection and missing value estimation algorithms for microarray analysis.

The linear discriminant analysis based on the generalized singular value decomposition (LDA/GSVD) has recently been introduced to circumvents the nonsingularity restriction that occur in the classical LDA so that a dimension reducing transformation can be effectively obtained for undersampled problems. We have established relationships between support vector machines (SVMs) and the generalized linear discriminant analysis applied to the support vectors. Based on the GSVD, the weight vector of the hard margin SVM is proved to be equivalent to the dimension reducing transformation vector generated by LDA/GSVD applied to the support vectors of the binary class. We have also shown that dimension reducing transformation vector and the weight vector of soft margin SVMs are related when a subset of support vectors are considered. These results can be generalized when kernelized SVMs and the kernelized KDA/GSVD are considered. Illustrating the relationship, it is shown that a classification problem can be interpreted as a data reduction problem.

To efficiently compute an updated solution without recomputing the solution from scratch every time when there is some minor changes in the data such as deletion and insertion of a few data points, we developed an adaptive KDA based on regularized MSE (KDA/RMSE). The regularization and the generalized singular value decomposition (GSVD) have been applied to overcome the nonsingularity restriction in Fisher's linear discriminant analysis for undersampled problems. For handling non-linearly separable cases, kernelized Fisher's discriminant analysis (KFD), kernel discriminant analysis based on GSVD (KDA/GSVD) and kernel discriminant analysis based on minimum squared error cost function (KDA/MSE) have been introduced.

By finding the relationship between classification and dimension reduction, we have also developed a method for gene selection. The identification of discriminant genes is one of the most fundamental steps in microarray gene expression data analysis to confirm recent discoveries in disease research or suggest new genes to be explored. We designed an L_1-norm support vector machines with recursive multiple feature elimination for gene selection in cancer classification, which is accurate and computationally efficient since it can remove several features at the same time for the identification of discriminant genes. We also introduced a gene selection method based on the cluster structure preserving dimension reduction method that has been used for feature extraction. Building a bridge between feature extraction and feature selection, a feature selection algorithm

is developed based on generalized linear discriminant analysis. Then, we successfully applied the feature selection algorithms to gene selection for cancer classification. The experimental results demonstrate that our method is capable of selecting linear and nonlinear features effectively so that competitive performance of classification can be obtained with linear classifiers in the dimension reduced space.

For missing value estimation in microarray analysis, we have designed a new scheme called LLSimpute. This method outperforms all other existing imputation methods on all the tests we have performed. Another related result is on gene selection in microarray analysis. We have also designed an efficient adaptive method for linear and nonlinear dimension reduction and classifier design.

Gene expression data often contain missing expression values. Effective missing value estimation methods are needed since many algorithms for gene expression data analysis require a complete matrix of gene array values. We have designed an imputation methods based on the least squares formulation are proposed to estimate missing values in the gene expression data, which exploit local similarity structures in the data as well as least squares optimization process. The proposed local least squares imputation method (LLSimpute) represents a target gene that has missing values as a linear combination of similar genes. The similar genes are chosen by k-nearest neighbors or k coherent genes that have large absolute values of Pearson correlation coefficients. Nonparametric missing values estimation method of LLSimpute are designed by introducing an automatic k-value estimator. In our experiments, the proposed LLSimpute method shows competitive results when compared with the other imputation methods for missing value estimation on various data sets and percentages of missing values in the data.

Another related problem of identification of discriminative genes for categorical phenotypes in microarray gene expression data analysis has been extensively studied especially for disease diagnosis. In the recent biological experiments, the continuous phenotypes have been also dealt with. For example, the extent of programmed cell death (apoptosis) can be measured by the level of Caspase3 enzyme. Thus, the effective gene selection method for continuous phenotypes is desirable. We designed a three-stage framework for gene expression data analysis based on L_1-norm support vector regression (L_1-SVR). The first stage ranks genes by recursive multiple feature elimination based on L_1-SVR. In the second stage, the minimal genes are determined by a kernel regression, which yield the lowest ten-fold cross validation error. In the last stage, the final nonlinear regression model is built with the minimal genes and optimal parameters found by leave-one-out cross validation. The experimental results show a significant improvement over the current state-of-art approach, i.e. the two-stage process, which consists of the gene selection based on L_1-SVR and the third stage of the proposed method.

Regularization and the generalized singular value decomposition have been applied to overcome the nonsingularity restriction in Fisher's linear discriminant analysis for undersampled problems when the dimension of the data space is higher than the number of

data points. Recently, kernelized nonlinear extensions of Fisher's discriminant analysis, discriminant analysis based on generalized singular value decomposition (LDA/GSVD), and discriminant analysis based on the minimum squared error formulation (MSE) have been introduced for handling undersampled problems and nonlinearly separable data sets. We have designed an efficient algorithm for adaptive linear and nonlinear kernel discriminant analysis based on regularized MSE, called adaptive KDA/RMSE. It can efficiently compute the updated and downdated solutions when data points are appended or removed.

Recently, kernelized nonlinear extensions of Fisher's discriminant analysis, discriminant analysis based on generalized singular value decomposition (LDA/GSVD), and discriminant analysis based on the minimum squared error formulation (MSE) have been introduced for handling undersampled problems and nonlinearly separable data sets. We have designed an efficient algorithm for adaptive linear and nonlinear kernel discriminant analysis based on regularized MSE, called adaptive KDA/RMSE. It can efficiently compute the updated and downdated solutions when data points are appended or removed. This adaptive classifier avoids the computationally expensive eigenvalue decomposition (EVD) updating, which would be necessary for designing an adaptive classifier for directly updating linear discriminant analysis. In adaptive KDE/RMSE, updating and downdating of the computationally expensive eigenvalue decomposition (EVD) or singular value decomposition (SVD) is approximated by updating and downdating of the QR decomposition achieving an order of magnitude speed up. This fast algorithm for adaptive kernelized discriminant analysis is designed by utilizing regularization and the relationship between linear and nonlinear discriminant analysis and the MSE. In addition, an efficient algorithm to compute leave-one-out cross validation is also introduced by utilizing downdating of KDA/RMSE.

We also proposed several multiclass classifiers based on generalized LDA algorithms, taking advantage of the dimension reducing transformation matrix without requiring additional training or any parameter optimization. A marginal linear discriminant classifier, a Bayesian linear discriminant classifier, and a one-dimensional Bayesian linear discriminant classifier are introduced for multiclass classification. Our experimental results illustrate that these classifiers produce higher ten-fold cross validation accuracy than kNN and centroid based classification in the reduced dimensional space providing efficient general multiclass classifiers.

Another significant theoretical discovery from our research is the relationship between support vector machines (SVMs) and the generalized linear discriminant analysis applied to the support vectors. Based on the GSVD, the weight vector of the hard-margin SVM is proved to be equivalent to the dimension reducing transformation vector generated by LDA/GSVD applied to the support vectors of the binary class. We also show that the dimension reducing transformation vector and the weight vector of soft-margin SVMs are related when a subset of support vectors are considered. These results can be generalized when kernelized SVMs and the kernelized LDA/GSVD called KDA/GSVD

are considered. Through these relationships, it is shown that support vector classification is related to data reduction as well as dimension reduction by LDA/GSVD.

Major Research and Education Activities

During the four years of the project (the first three initial years and one year co-cost extension) we have designed several linear and nonlinear dimension reduction methods in conjunction with the efficient classification methods which can be effectively applied to several protein structure prediction related problems. Two of the major research activities were focused on designing the efficient algorithms and obtaining substantial implementation results on very large data sets on protein structure. In addition, we have designed new algorithms for data analysis in microarray analysis including the missing value estimation, gene selection, and clustering for cancer class discovery.

The prediction of protein secondary structure is an important step in the prediction of protein tertiary structure. We have developed a new protein secondary structure prediction method SVMpsi to improve the current level of prediction by incorporating new tertiary classifiers and their jury decision system, and the PSI-BLAST PSSM profiles. Additionally, efficient methods to handle unbalanced data and a new optimization strategy for maximizing the Q_3 measure are developed. The SVMpsi produces the highest published Q_3 and SOV94 scores on both the RS126 and CB513 data sets to date. For a new KP480 set that we prepared for testing protein secondary structure prediction methods, the prediction accuracy of SVMpsi was $Q_3 = 78.5\%$ and SOV94 $= 82.8\%$. The SVMpsi results in CASP5 competition illustrate that it is another competitive method to predict protein secondary structure.

Another related problem that we explored is the prediction of protein relative solvent accessibility. It gives us helpful information for the prediction of tertiary structure of a protein. The SVMpsi method which uses support vector machines (SVMs) and the position specific scoring matrix (PSSM) generated from PSI-BLAST has been applied to achieve better prediction accuracy of the relative solvent accessibility. We have introduced a three dimensional local descriptor which contains information about the expected remote contacts by the long-range interaction matrix as well as neighbor sequences. Moreover, we applied feature weights to kernels in support vector machines in order to consider the degree of significance that depends on the distance from the specific amino acid. The highest prediction accuracies for relative solvent accessibility based on a two state-model, for 25%, 16%, 5%, and 0% accessibility, and three state prediction results of 64.5% accuracy with 9%;36% threshold have been obtained. The support vector machine approach has successfully been applied for solvent accessibility prediction by considering long-range interaction and handling unbalanced data.

In microarray analysis, missing value estimation and multiple gene selection, and clustering methods for cancer class discovery have been studied. In the process, adaptive linear and nonlinear dimension reduction and adaptive classifier design schemes have been designed. They are playing very important roles in designing fast and effective dimension reduction methods in protein structure prediction. The major findings resulting from the research activities so far include the discovery of optimal encoding

schemes for amino acids for protein structure prediction and efficient implementation of the support vector machine classifier utilizing its relationship to the linear discriminant analysis. In addition, a missing value estimation scheme called LLSimpute was designed. This method outperforms all other existing imputation methods on all the tests we have performed. Another related result is on gene selection in microarray analysis. We have also designed an efficient adaptive method for linear and nonlinear dimension reduction and classifier design.

The data sets in the problems that we consider typically have the characteristics of being undersampled, i.e., the number of data items is much smaller than the data dimension. Due to this fact, some novel algorithms had to be designed for data analysis since some traditional statistical methods assume the availability of enough data points and this cannot be satisfied in many biological data sets such as microarray data sets due to high cost associated with data collection. We designed the linear discriminant analysis based on the generalized singular value decomposition (LDA/GSVD) earlier to circumvents the nonsingularity restriction that occur in the classical LDA so that a dimension reducing transformation can be effectively obtained for undersampled problems. We have established relationships between support vector machines (SVMs) and the generalized linear discriminant analysis applied to the support vectors. Based on the GSVD, the weight vector of the hard margin SVM is proved to be equivalent to the dimension reducing transformation vector generated by LDA/GSVD applied to the support vectors of the binary class. These results can be generalized when kernelized SVMs and the kernelized KDA/GSVD are considered.

To efficiently compute an updated solution without recomputing the solution from scratch every time when there is some minor changes in the data such as deletion and insertion of a few data points, we developed an adaptive KDA based on regularized MSE (KDA/RMSE). The regularization and the generalized singular value decomposition (GSVD) have been applied to overcome the nonsingularity restriction in Fisher's linear discriminant analysis for undersampled problems. For handling non-linearly separable cases, kernelized Fisher's discriminant analysis (KFD), kernel discriminant analysis based on GSVD (KDA/GSVD) and kernel discriminant analysis based on minimum squared error cost function (KDA/MSE) have been introduced.

Gene expression data often contain missing expression values. Effective missing value estimation methods are needed since many algorithms for gene expression data analysis require a complete matrix of gene array values. We have designed an imputation method based on the least squares formulation are proposed to estimate missing values in the gene expression data, which exploit local similarity structures in the data as well as least squares optimization process. The proposed local least squares imputation method (LLSimpute) represents a target gene that has missing values as a linear combination of similar genes. The similar genes are chosen by k-nearest neighbors or k coherent genes that have large absolute values of Pearson correlation coefficients. Nonparametric missing values estimation method of LLSimpute are designed by introducing an auto-

matic k-value estimator. In our experiments, the proposed LLSimpute method shows competitive results when compared with the other imputation methods for missing value estimation on various data sets and percentages of missing values in the data.

Another related problem of identification of discriminative genes for categorical phenotypes in microarray gene expression data analysis has been extensively studied especially for disease diagnosis. In the recent biological experiments, the continuous phenotypes have been also dealt with. For example, the extent of programmed cell death (apoptosis) can be measured by the level of Caspase3 enzyme. Thus, the effective gene selection method for continuous phenotypes is desirable. We designed a three-stage framework for gene expression data analysis based on L_1-norm support vector regression (L_1-SVR). The first stage ranks genes by recursive multiple feature elimination based on L_1-SVR. In the second stage, the minimal genes are determined by a kernel regression. In the last stage, the final nonlinear regression model is built with the minimal genes and optimal parameters found by leave-one-out cross validation. The experimental results show a significant improvement over the current state-of-art approach, i.e. the two-stage process, which consists of the gene selection based on L_1-SVR and the third stage of the proposed method.

Recently, kernelized nonlinear extensions of Fisher's discriminant analysis, discriminant analysis based on generalized singular value decomposition (LDA/GSVD), and discriminant analysis based on the minimum squared error formulation (MSE) have been introduced for handling undersampled problems and nonlinearly separable data sets. We have designed an efficient algorithm for adaptive linear and nonlinear kernel discriminant analysis based on regularized MSE, called adaptive KDA/RMSE. It can efficiently compute the updated and downdated solutions when data points are appended or removed. This adaptive classifier avoids the computationally expensive eigenvalue decomposition (EVD) updating, which would be necessary for designing an adaptive classifier for directly updating linear discriminant analysis. In adaptive KDE/RMSE, updating and downdating of the computationally expensive eigenvalue decomposition (EVD) or singular value decomposition (SVD) is approximated by updating and downdating of the QR decomposition achieving an order of magnitude speed up. This fast algorithm for adaptive kernelized discriminant analysis is designed by utilizing regularization and the relationship between linear and nonlinear discriminant analysis and the MSE. In addition, an efficient algorithm to compute leave-one-out cross validation is also introduced by utilizing downdating of KDA/RMSE.

The educational activities include the development of advance course material at University of Minnesota "CSci 8323: Solving Linear Least Squares Problems". In the course, numerical techniques for data analysis and feature extraction are discussed using the framework of matrix rank reduction. The problems from the bioinformatics as well as the support vector machine classification techniques are newly incorporated in the class. In addition, several imputation schemes for missing value estimation in microarray analysis based on SVD and least squares formulation are introduced. The problems

from the bioinformatics as well as the support vector machine classification techniques are newly incorporated in the class. A Master's degree project created based on the presented material in the course has been successfully completed by a MS student. Two Ph.D. students who participated in the project have completed their Ph.D. degrees. Several related papers are published in refereed journals or submitted and currently under review.

After moving to the Georgia Institute of Technology, I have lead and organized a weekly seminar series in Computational Science and Engineering under the theme of data analysis and mining. The seminar served the role of providing a forum in connecting and exchanging research ideas/activities including those in data analysis and biology. In the seminar, my postdoctoral researcher and myself introduced our results on microarray analysis such as gene selection, missing value estimation, and ovarian cancer discovery results. In addition, several imputation schemes for missing value estimation in microarray analysis based on SVD and least squares formulation are introduced. The problems from the bioinformatics as well as the support vector machine classification techniques are newly incorporated in the class. The plan is to eventually organize the seminar series into a high level graduate course for the students in bioinformatics and computational biology.